PatchCleanser: Certifiably Robust Defense against Adversarial Patches for Any Image Classifier

Chong Xiang, Saeed Mahloujifar, Prateek Mittal

Princeton University

PRINCETON UNIVERSITY

PatchCleanser: Certifiably Robust Defense against Adversarial Patches for Any Image Classifier

Chong Xiang, **Saeed Mahloujifar, Prateek Mittal**

Princeton University

PRINCETON UNIVERSITY

PatchCleanser: Certifiably Robust Defense against Adversarial Patches for Any Image Classifier

Chong Xiang, Saeed Mahloujifar, Prateek Mittal

Princeton University

Adversarial Patch Attack: A Variant of Adversarial Examples

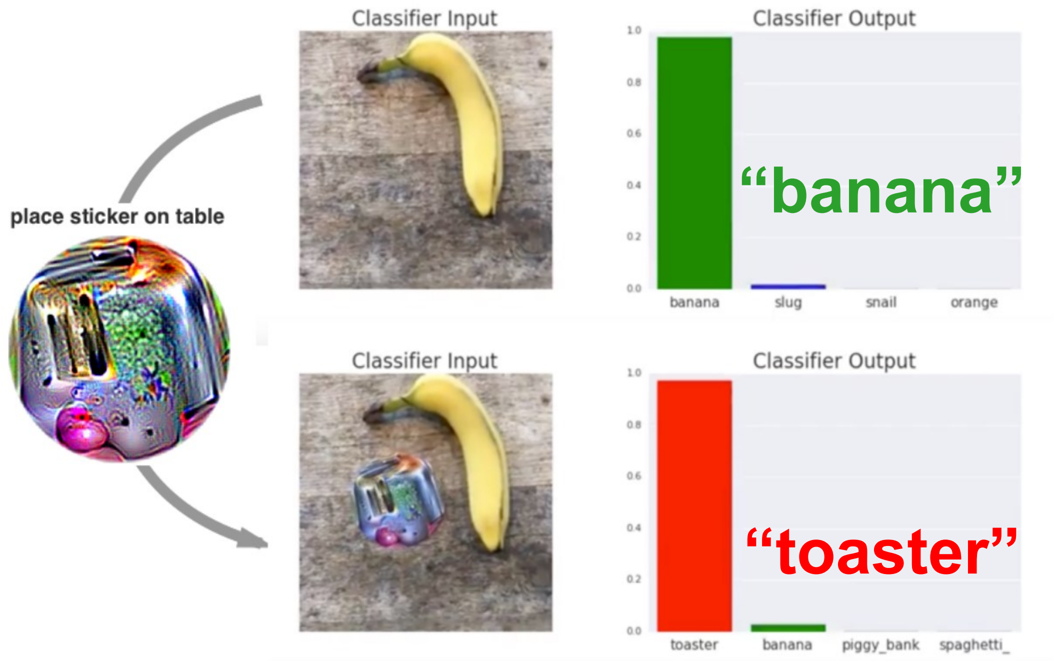
- All adversarial pixels within one local region (patch)
- Optimize the patch content for test-time model misclassification
- Print and attach the patch to the physical scene -- **a threat in the physical world!**

"Tiger cat"

Karmon et al. [ICML 2018]

"Speed Limit 80"

Yakura et al. [AAAI 2020]



Brown et al. [NeurIPS W 2017]

How can we build robust models against adversarial patches?

How to Quantify and Evaluate Robustness?

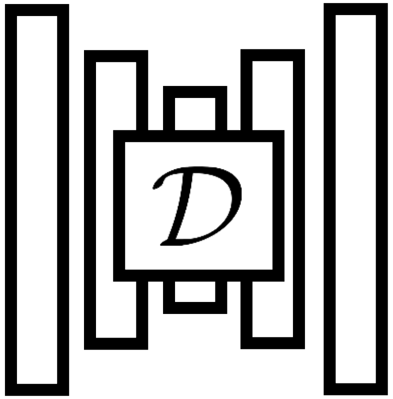
- Usually, people use a specific attack for robustness evaluation
- Problem: robustness evaluated today might be compromised by smarter adaptive attackers in the future

Evading Adversarial Example Detection Defenses with Orthogonal Projected Gradient Descent

Oliver Bryniarski, Nabeel Hingun, Pedro Pachuca, Vincent Wang, Nicholas Carlini

- Can we design defenses in a special way such that we can prove their robustness against any future adaptive attack strategies?
  - **Certifiable Robustness!**

Certificable Robustness: Formulation



Defense Model

Input Image
w/ ground-truth label

Patch Threat Model
(patch sizes, shapes, and location set)

Robustness Certificate

The model prediction is *always* “dog”, no matter what a **white-box adaptive** attacker within the threat model does

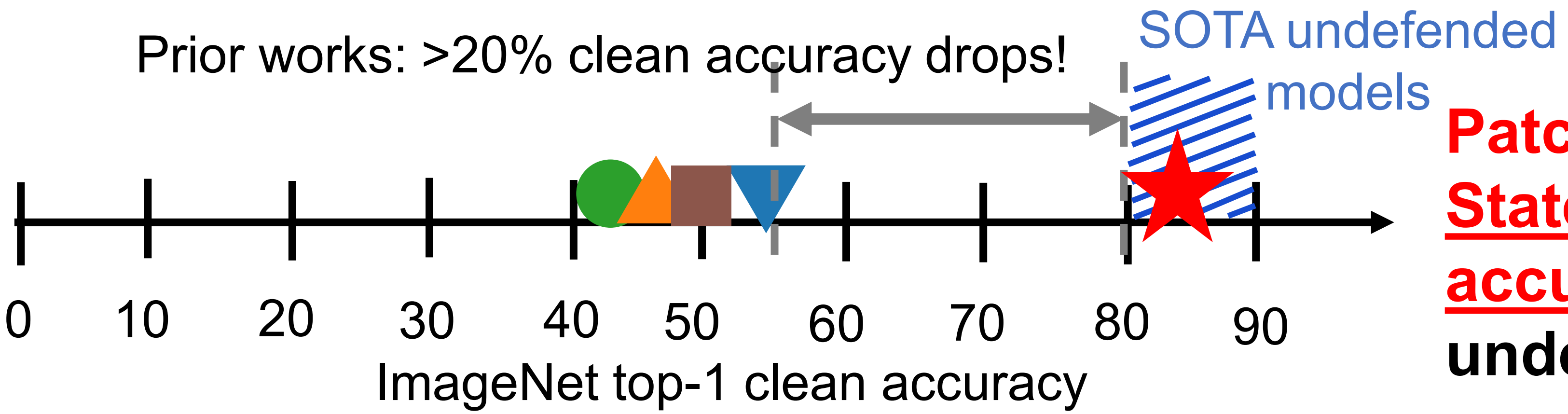
A typical patch threat model:
One 2%-pixel square patch with any content at any image location

Any patch content

Any patch location

Highlights of PatchCleanser

- **State-of-the-art certifiable robustness against adversarial patches**
  - Strong robustness guarantees!
- **A minimal cost of clean performance (accuracy without attack)**

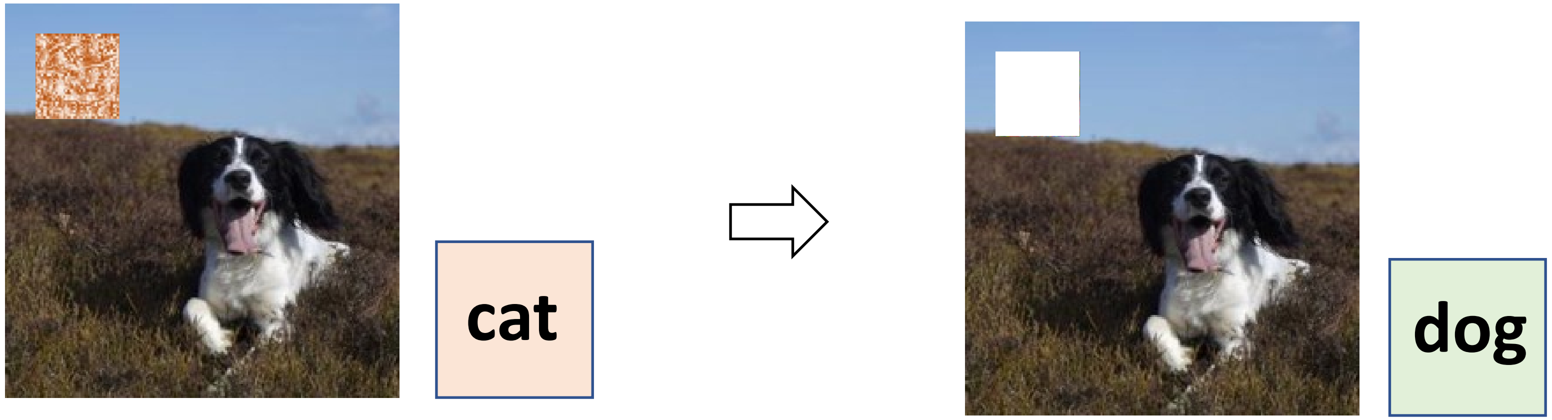


PatchCleanser: <u>State-of-the-art clean accuracy</u> even among undefended models!

- **The first defense with state-of-the-art certifiable robustness and clean performance**

PatchCleanser: A Pixel-Masking Defense

- Mask out the entire patch to neutralize adversarial effects



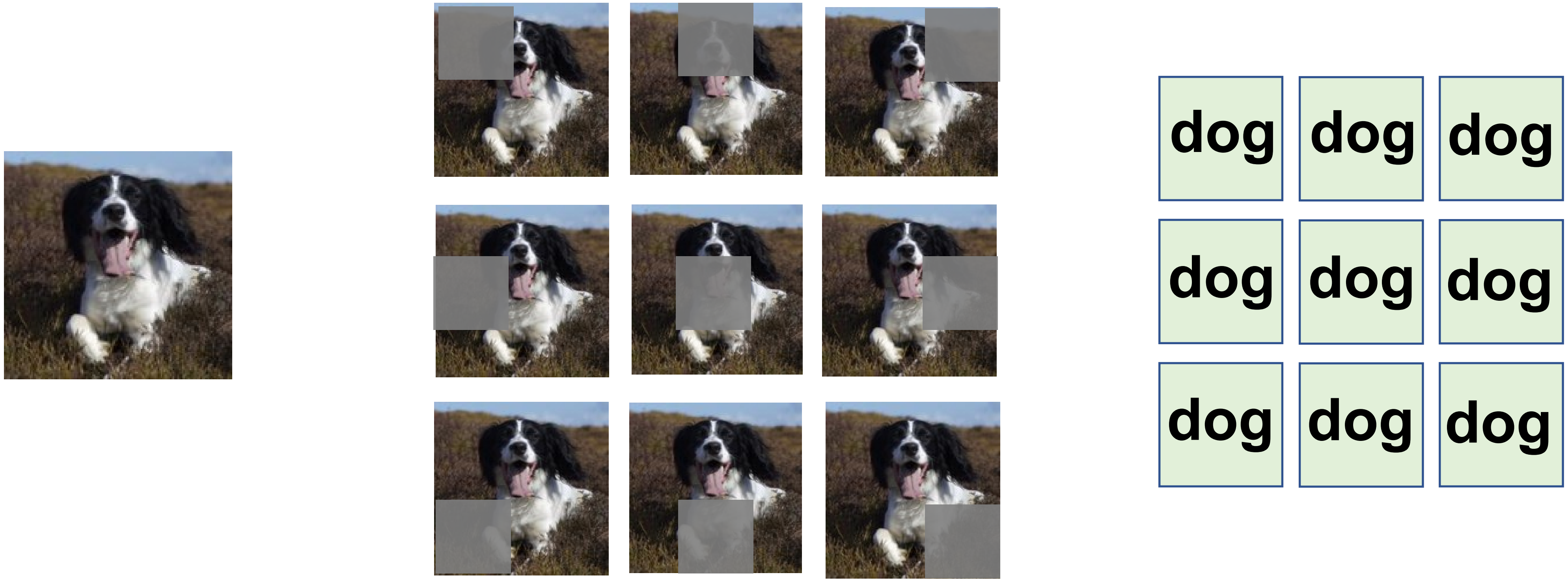
- Recover correct predictions using any state-of-the-art classifier

How to mask out the patch?

(in a certifiably robust manner)

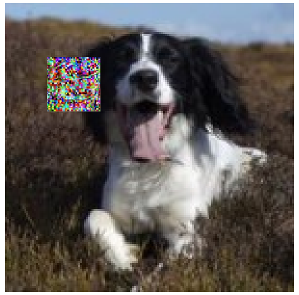
Intuition 1: Applying Small Masks to Clean Images Barely Changes Model Predictions

- We can still recognize the dog even with a small mask on the image



Intuition 2: Applying Small Masks to Adversarial Images Can Change Model Predictions

- When we mask out the patch, we can get the correct prediction label back

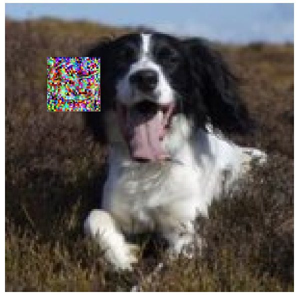


Focus on one patch
(can be extended to multiple patches)

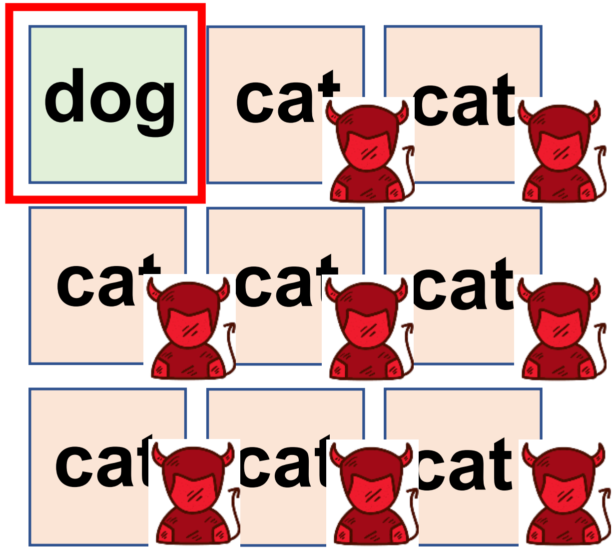
dog	cat	cat
cat	cat	cat
cat	cat	cat

Question: How Can We Settle This Disagreement?

- How to identify the correct prediction label?



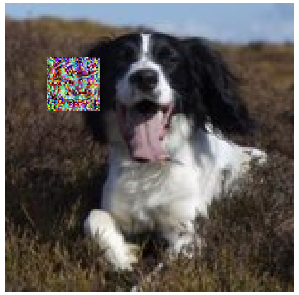
Output the disagreer?



What if the attacker introduces other prediction labels?

Question: How Can We Settle This Disagreement?

- How to identify the correct prediction label?



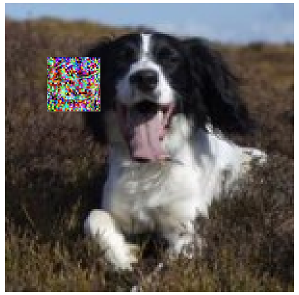
Output the disagreer?

dog	cat	cat
cat	cat	cat
cat	cat	fox

What if the attacker introduces other prediction labels?

Question: How Can We Settle This Disagreement?

- How to identify the correct prediction label?



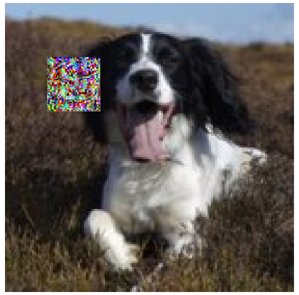
Output the disagreer?

dog	cat	cat
cat	cat	cat
cat	fox	cat

What if the attacker introduces other prediction labels?

Question: How Can We Settle This Disagreement?

- How to identify the correct prediction label?



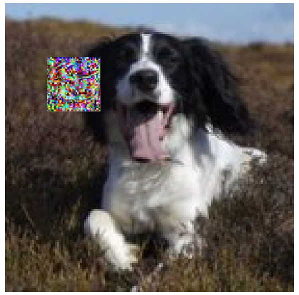
Output the disagreer?

dog	cat	cat
cat	cat	cat
fox	cat	cat

What if the attacker introduces other prediction labels?

Question: How Can We Settle This Disagreement?

- How to identify the correct prediction label?



Output the disagreer?

dog	cat	cat
cat	cat	cat
fox	cat	cat

How can we distinguish <u>patch-removing</u> masks from other masks?

How can we distinguish between “dog” and “fox”?

Add a Second Mask!

- Analyze model predictions on images with two masks
- To determine if the first mask removes the patch or not

Case 1: the First Mask Removes the Patch

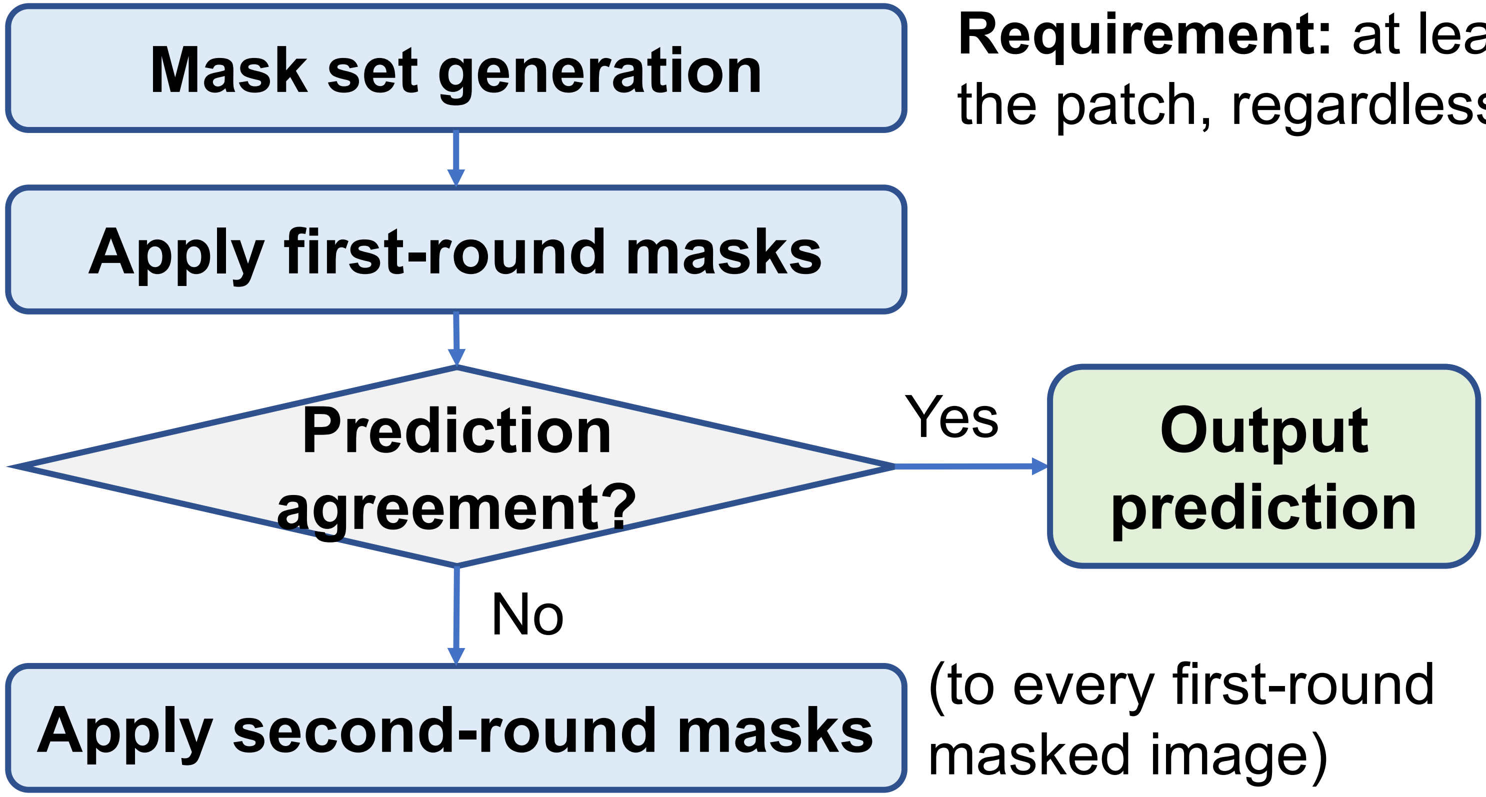
- The second mask is applied to a *clean* image
- Two-mask predictions reach a unanimous *agreement*

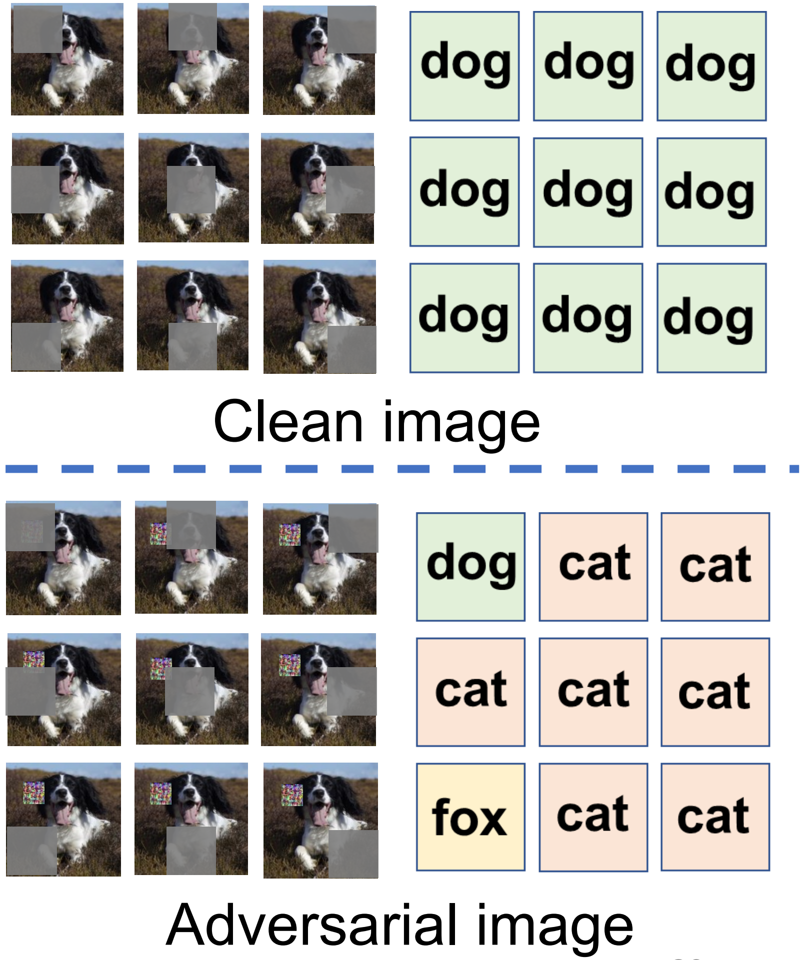
Case 2: the First Mask Does not Remove the Patch

- The second mask is applied to an *adversarial* image
- Two-mask predictions have *disagreement*

Double-masking: Defense via Two Rounds of Masking

Requirement: at least one mask can remove the patch, regardless of the patch location





Double-masking: Defense via Two Rounds of Masking

Mask set generation

Apply first-round masks

Prediction agreement?

Yes → **Output prediction**

No

Apply second-round masks

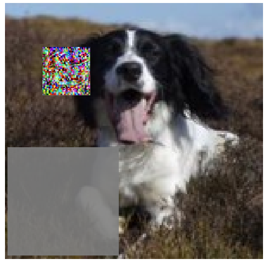
(to every first-round masked image)

Prediction agreement?

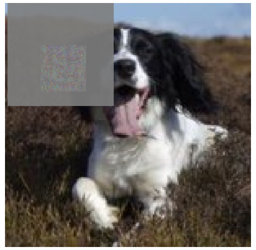
Yes → **Output prediction**

No

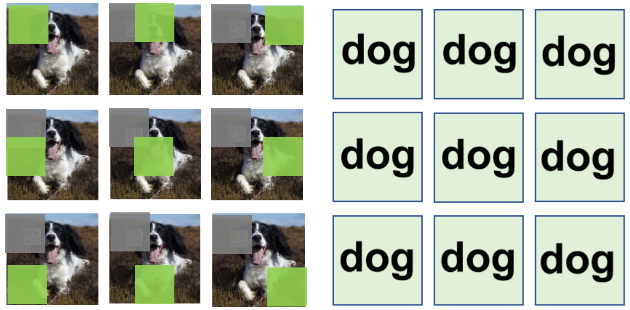
move on to next masked image



Adversarial pixels left

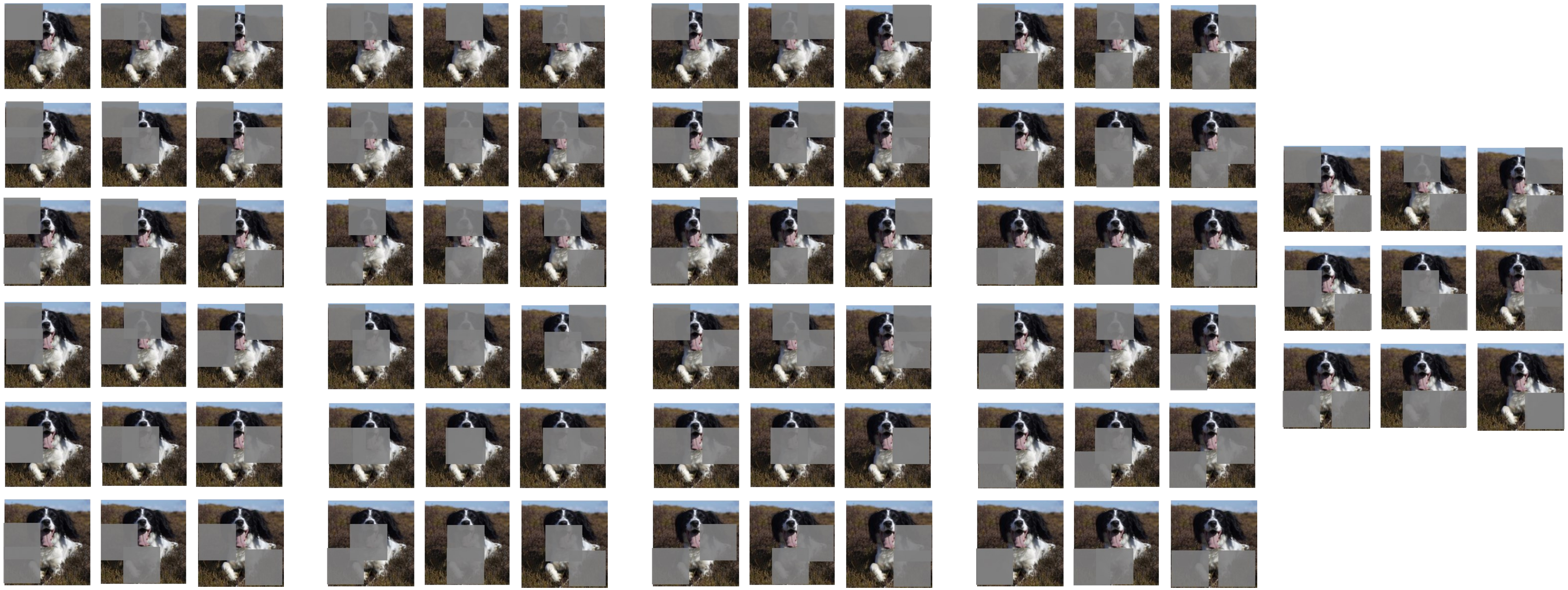


the first mask removes the patch!

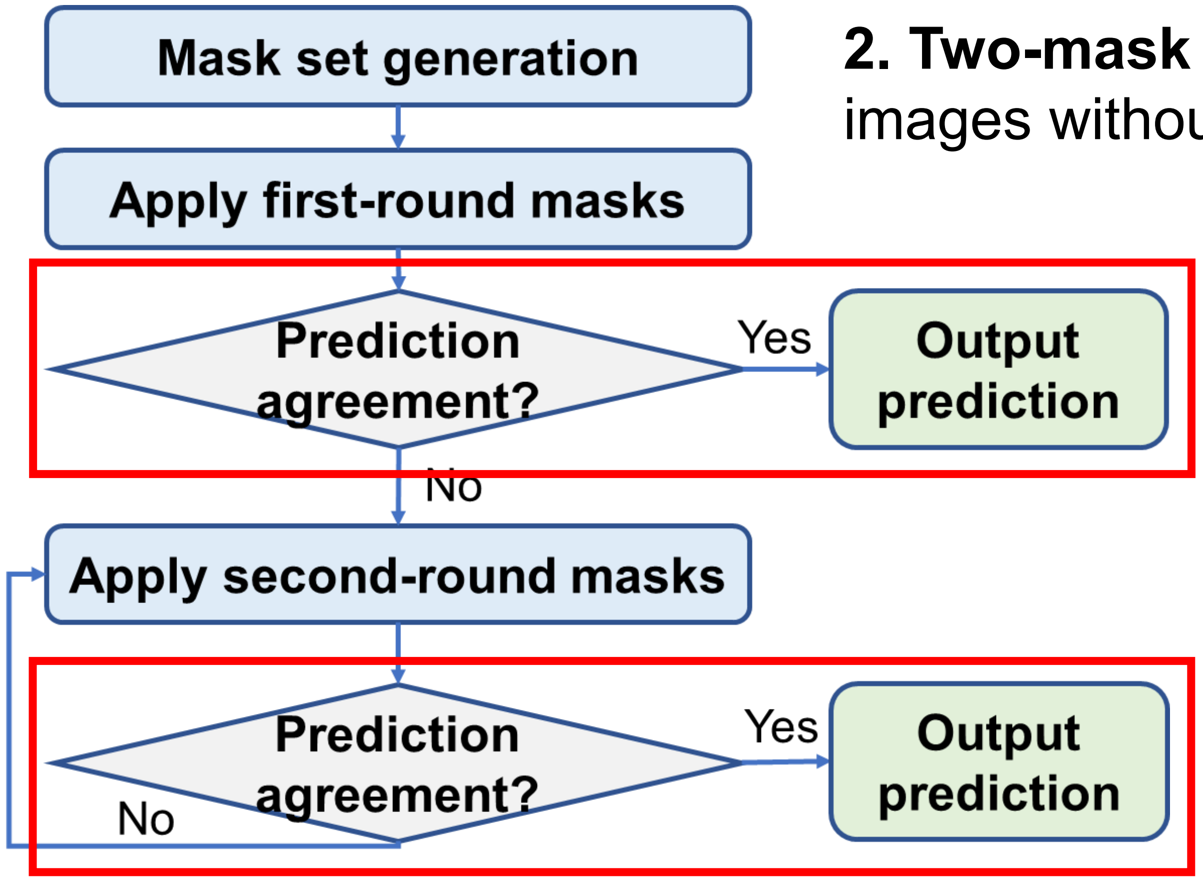


Robustness Certification

- **Two-mask correctness** implies certifiable robustness
  - Model predictions on all possible two-masked images are correct



Proof (No Math Needed): Never Return Incorrect Labels



1. Mask set: at least one mask can remove the patch
2. Two-mask correctness: predictions on masked images without adversarial pixels are all correct

One-mask prediction

1. At least one correct one-mask prediction
   - A first-round mask removes the patch
2. Enforce disagreement with other labels (if any)
3. Never returns incorrect labels

Two-mask prediction

1. At least one correct two-mask prediction
   - A second-round mask removes the patch
2. Enforce disagreement with other labels (if any)
3. Never returns incorrect labels

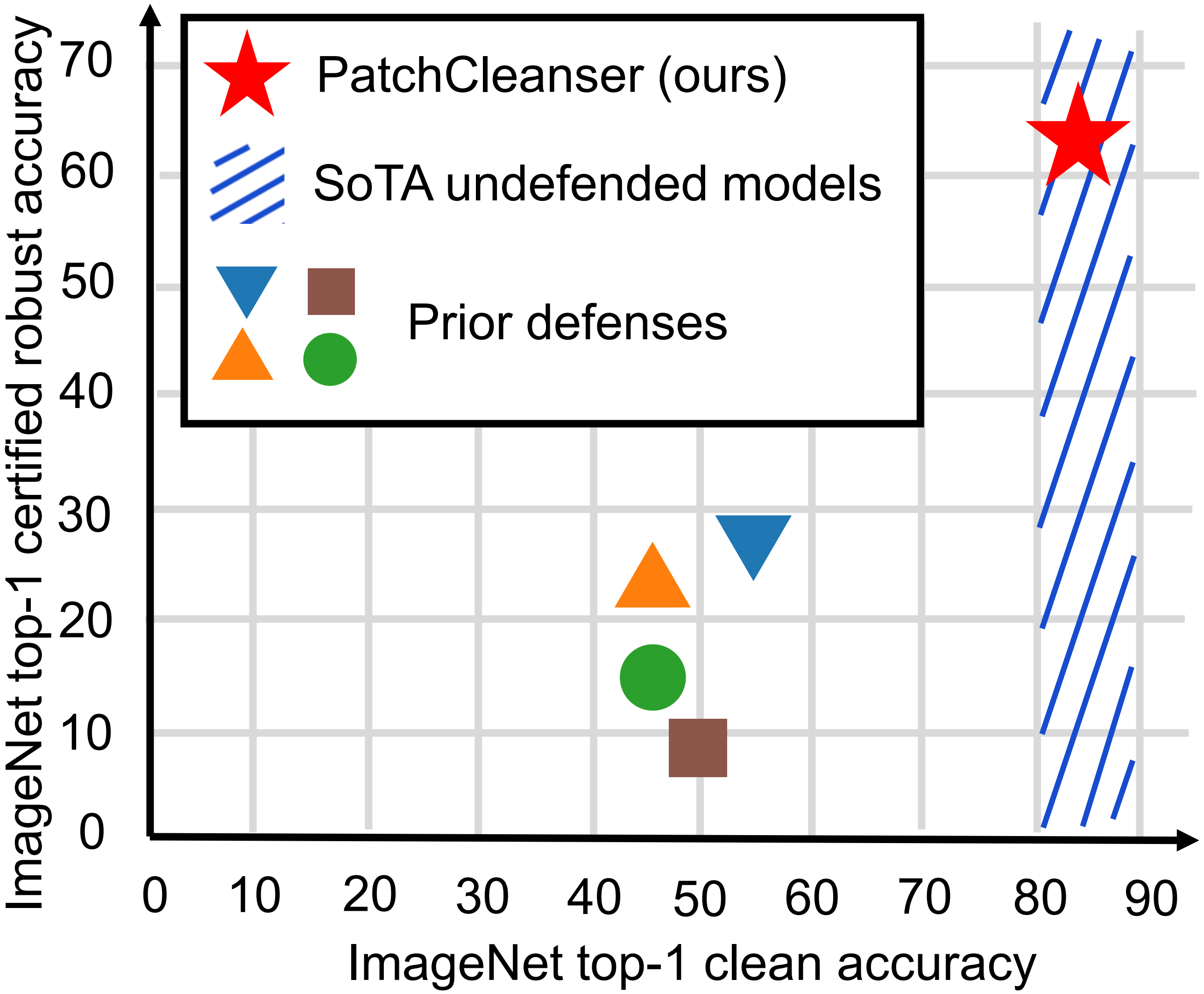
Evaluation Setup

- **Clean accuracy**
  - Fraction of correctly classified test images

- **Certified robust accuracy**
  - Fraction of test images we can certify the robustness for
  - i.e., two-mask correctness

PatchCleanser Performance

- **ImageNet** evaluation: robustness evaluated for a 2%-pixel square patch anywhere on the image
- PatchCleanser's **clean accuracy** (83.9%) falls within the range of state-of-the-art undefended models (~1% accuracy drops)
- PatchCleanser's **certified robust accuracy** (62.1%) is even higher than clean accuracy of prior works



Takeaways

- **PatchCleanser**
  - pixel masking defense
  - certifiable robustness for recovering correct prediction labels
- **The first certified defense with 83+% accuracy on ImageNet**
  - As well as state-of-the-art certifiable robustness
- **Compatible with any state-of-the-art image classifiers**
  - While prior works all rely on specific model architectures (e.g., small receptive fields)

ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

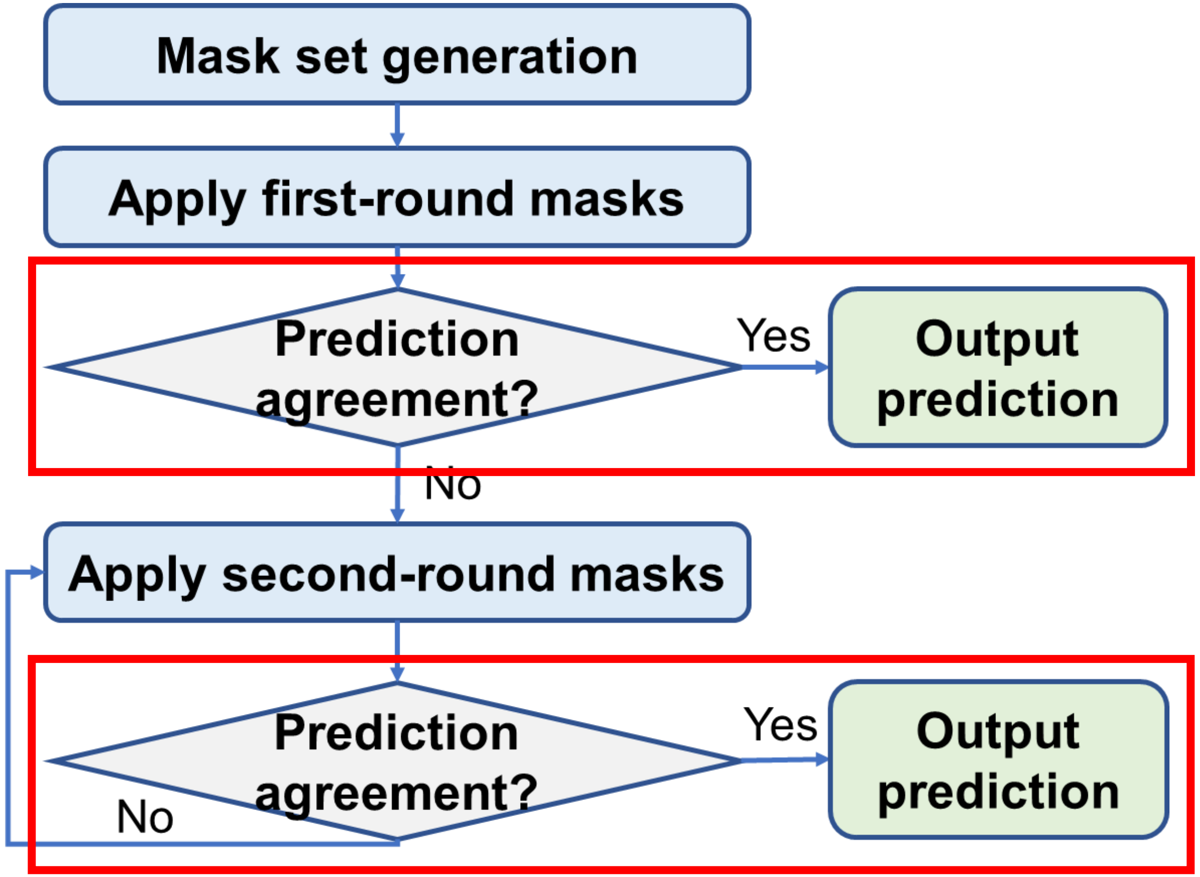
ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED

artifact

leaderboard

paper list

Backup Slide: Conservative in Returning Incorrect Labels on Clean Images

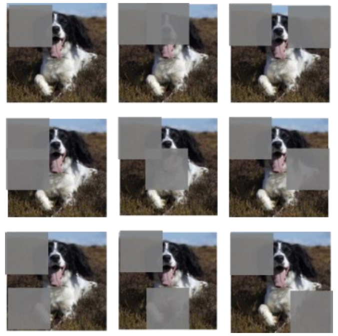


Return incorrect labels when:

1. One-mask predictions agree on incorrect labels

cat	cat	cat
cat	cat	cat
cat	cat	cat

2. Two-mask predictions agree on incorrect labels



cat	cat	cat
cat	cat	cat
cat	cat	cat

Rarely happens in the clean setting!

Backup Slide: Mask Set

- **Requirement:** at least one “mask” can remove all adversarial pixels
- **Multiple patches**

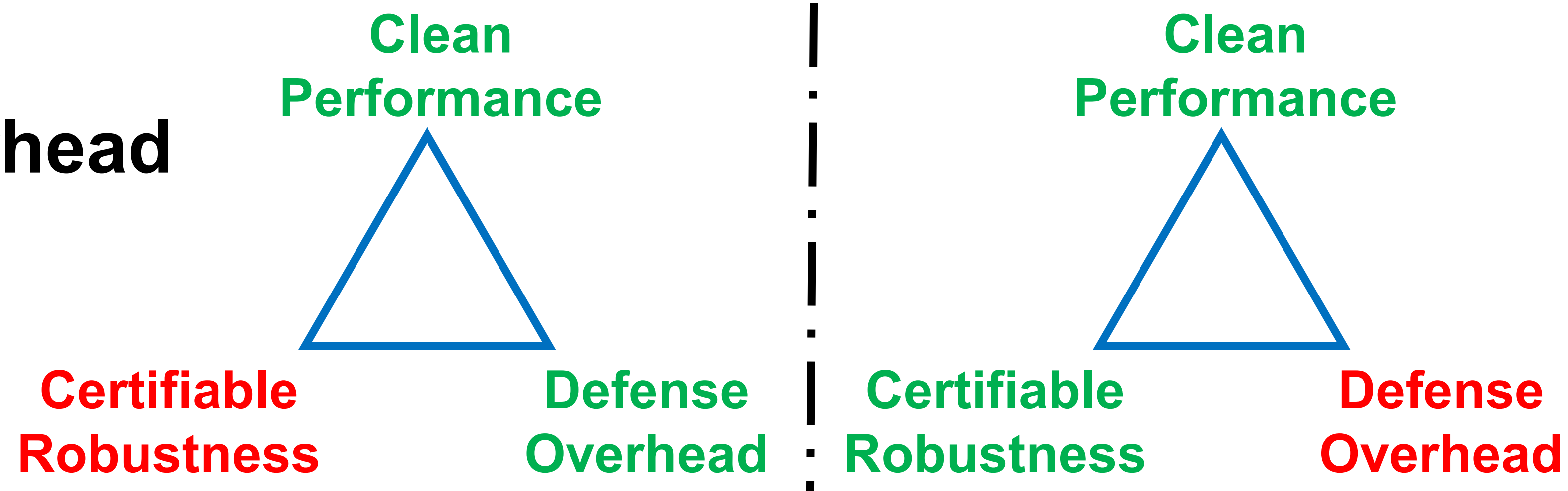
	Clean accuracy	Certified robust accuracy
two 1%-pixel squares	83.8%	45.8%
one 2%-pixel square	83.8%	63.2%

- **Different patch shapes**

	Clean accuracy	Certified robust accuracy
Any 1%-pixel rectangle	85.4%	49.8%
Any 1%-pixel square	84.2%	68.2%

Backup Slide: Limitation

- **Requires additional defense parameters for mask set generation**
  - An insecure mask set undetermined the robustness
- **Trade-off between robustness and overhead**
  - Requires evaluating predictions on multiple masked images



Undefended models

1. Good clean performance
2. Zero robustness
3. Good defense overhead

PatchCleanser

1. Good clean performance
2. Good certifiable robustness
3. Poor defense overhead

Thank you!

Chong Xiang
Princeton University
cxiang@princeton.edu

Saeed Mahloujifar
Princeton University
sfar@princeton.edu

Prateek Mittal
Princeton University
pmittal@princeton.edu

Technical Report

GitHub